Enhanced Performance of Search Engine with Multitype Feature Co-Selection of K-Means Clustering Algorithm

K.Parimala
Assistant Professor,
MCA Department,
NMS.S.Vellaichamy Nadar College, Madurai

Dr.V.Palanisamy,
Professor & Head of the Department
Department of CSE,
Alagappa University, Karaikudi

Abstract

Information world meet many confronts nowadays and one such, is data retrieval from a multidimensional and heterogeneous data set. Han & et al carried out a trail for the mentioned challenge. A novel feature co-selection for web document clustering is proposed by them, which is called Multitype Features Co-selection for Clustering (MFCC). MFCC uses intermediate clustering results in one type of feature space to help the selection in other types of feature spaces. It reduces effectively of the noise introduced by "pseudoclass" and further improves clustering performance. This efficiency also can be used in data retrieval, by implementing the MFCC algorithm in ranking algorithm of Search Engine technique. The proposed work is to apply the MFCC algorithm in search engine architecture. Such that the information retrieves from the dataset is retrieved effectively and shows the relevant retrieval.

Keywords: MFCC algorithm, Search Engine, Ranking algorithm, Information Retrieval.

I. INTRODUCTION

Information is being created and it is becoming available in quantities by the log on possibilities proliferate. There is a great deal of excitement about the electronic information superhighway that enables information seekers to access the diverse and large information sources. However, the realization of making information available to users almost straight away, commonly referred to as, the 'information explosion', is already becoming a mixed blessing without better methods to filter, retrieve and manage this potentially unlimited influx of information. Users face 'information overload' and they require tools to explore the vast universe of information [1].

The information seeking behavior of a user depends on education, access to library and the length of the time to devote for information seeking. Naturally, most individuals seek information from friends, neighbors, colleagues and libraries among others. With the advent of internet, Many Professionals, Researchers and highly placed individuals seek information from the internet now. Information retrieval is concerned with the explanation of the information and other contents of documents. The establishments of various large databases, which are mounted on computers, are made available to anyone in the world. It has a significant impact on the effectiveness and efficiency of the retrieval of information.

The field of information retrieval has continued to change and grow, Collection have become larger, Computers have become more powerful, Broadband and mobile internet is widely assumed, Complex interactive search can be done on home computers or mobile devices, and so on. Furthermore, as large-scale commercial search companies find new enormous ways to exploit the user data they collect.

IR evaluation [2] is challenged by variety and fragmentation in many respects, diverse tasks and metrics, Heterogeneous collections, Different systems, Alternative approaches for managing the experimental data. Evaluation of using large data sets is often desirable in IR in order to provide corroboration of claimed improvements in search effectiveness and search efficiency, sometimes as well as both in nature.

Search in the real world is highly inherently interactive in the real world. A search is a non-trivial search task consists of stages with different sub-goals and specific search tactics. Search systems are becoming more complicated and are presenting richer results for example, combinations of documents, images, and videos. Simple summaries are no longer sufficient for emerging application areas.

Search engines [3] are powerful intellectual technologies that structure people's thinking and activities. The presumption that a general purpose search engine can fulfill all needs of a specific site, a specific user group, or a specific collection without parameter tuning is wrong. Search as encountered in its most general mode on the web is highly effective and convenient for a majority of search transactions. However, for the numerous specific needs and tasks in various organizations.

The information seeking can be a cumbersome process which is only partially supported: multilingual and cross-cultural issues, quality assurance requirements, in house jargon, etc., Interact to make site-specific and adaptable search technology a necessity. Since users expect similar convenience and effectiveness from in house system nowadays, that they are used to in a web context. Many Organizations outsource their search their needs to web search site-level indexes. In practice however, a tailored enterprise search solution would be more effective, if not too costly.

Search engines have conditioned users to interact with information in ways that are suboptimal for many types of search tasks and for deeper learning. While the convenience of contemporary search engines enables fast, easy and efficient access to certain types of information, the search behaviors learned through interactions. When translated in to tasks where deeper learning is required, often fail, search engines are currently optimized for look-up tasks and not tasks that require more sustained interactions with information

The challenge is to develop architecture for information access that can ensure freshness and coverage of information in a rapidly growing web. It is especially challenging to maintain freshness and coverage in a centralized search engine. The current approach is to visit frequencies for different types of pages or websites. There is something inherently wrong with waiting for a Google crawler to come around and pick up new content before it can be "found" by people and as the web grows the issues of freshness will get worse.

The proposed work is to find solutions for the challenges that were discussed in IR and in Web Service. MFCC algorithm is implemented to rectify those challenges, it have proved its efficiency in clustering of heterogeneous and multidimensional data from the data sets. An architecture is designed to solve the above said challenges such as search and retrieve, fresh pages retrieval from the data set as new contents were add up, time to retrieve in effective and efficient manner.

II. PROPOSED WORK

IR systems play a central role in helping people to develop their search skills, Also in supporting a larger variety of more sophisticated search strategies, and in supporting deeper learning experiences through the provision of integrative work environments that include a variety of tools for exploring information and a variety of interfaces that support different types of information behaviors, interactions and outcomes. Search with task and person context requirement follows as, Novel mixture of search and recommendation methods, Novel retrieval models, and Evaluation methods.

The feature selection plays a vital role in machine learning, data mining, information retrieval, etc. The goal of feature selection is to identify those features relevant to achieve a predefined task. Many researchers have been to find how to search feature space and evaluate them.

Multi-type Features Co-Selection for Clustering (MFCC), is an algorithm to exploit heterogeneous features of a web page like URL, anchor text, hyperlink etc., and to find discriminated features for unsupervised learning [4]. The additional information is to enhance the feature selection in other spaces. Consequently, the better feature set co-selected by heterogeneous features will produce better clusters in each space. After that, the better intermediate result will further improve co-selection in the next iteration. Finally, feature co-selection is executed iteratively and can be well integrated into an iterative clustering algorithm.

MFCC find more discriminative features and improve clustering performance. Hard k-means clustering version is chosen as the basic algorithm. [5]

The architecture of searching is designed as follows (refer Fig-1): (i) the search word or keyword is validated (vsm model) and a dictionary list is created, (ii) the data set or the data base or user group is classified in to feature spaces, (iii) the data set is validated (vsm model), (iv) the data set is validated and feature spaced according to the keyword, (v) best is chosen in each feature space using, the feature selection score(FSS), (vi) then the features (SF) are co-selected using ranking formula to produce the final rank result.

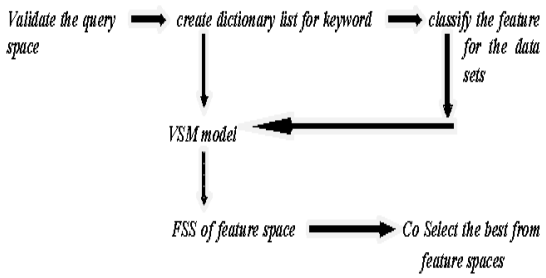

Fig-1 Data flow of MFCC

As Pseudoclass was introduced to the class identifier such as text, structure, utility, etc. are removed and clusters into feature spaces. Iterative feature clustering helps to remove outliers, so that the problem of fresh or new web pages in search results is also solved.

MFCC has proved its clustering efficiency in web documentation for the databases like www.opendirectory, www.project.com. The result shows that the clustering features have better relevancy than any other. Also it has provided its integrity in text classifiers also.

MFCC is better than the ranking algorithm. Since ranking algorithm, prepares the rank list based on the relevancy score. Then, links are matched according to the citations and grouped. But in MFCC it groups or classifies the data set in to feature spaces. In that, the feature selection score (FSS) is calculated using the statistical formulae: Information Gain (IG), Chi-Square, Correlation Coefficient (CC) and GSS Coefficient (GSS) in each feature spaces. Then using ranking formulae, SF, best data's are co-selected among the feature spaces. This is clustered iteratively.

MFCC trains the noisy data and uses that also for the score, no such facility in ranking algorithm. Such reliability can be executed in search engine technology to improve the ranking results.

The proposed architecture is likely to employ in the database index shown in Fig-2

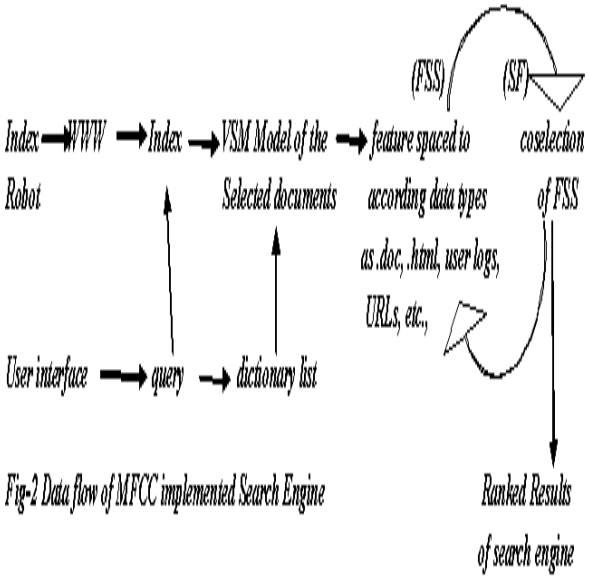

Fig-2 Data flow of MFCC implemented Search Engine

III. Results

The evaluation approach measured the quality of generated clusters by comparing them with a set of categories created manually. It performed in a test data set. The test data set contains 255 articles evenly classified in to at least 10 feature spaces (Table-1). In the experiments, MFCC algorithm ran a test on categories having highest number of documents.

MFCC algorithm clusters the data set according to the query term or search key. TF-IDF is calculated and the following result is for Chi-Square, Correlation Co-efficient, GSS Co-efficient and Information Gain for each feature class (Fig-3). The similarity of objects is the cosine of vectors in VSM model. TF-IDF with "iterative feature clustering" scheme was used to calculate the weight of each vector dimension.

Classes	No of documents	Related terms	Total term frequency
ASP	2	22	23
CSS	10	1439	6771
Gif	144	975	976
Html	25	14210	63392
Jpeg	18	3554	3659
Js	19	4935	38415
Pdf	5	249229	398036
Php	10	1670	4644
Png	9	245	245
Ppt	13	193505	208541

Table-1 Feature classes of test dataset.

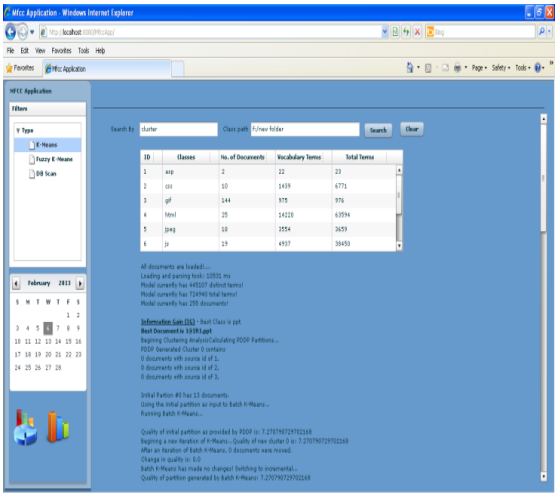

Fig-3 MFCC Search results

MFCC search architecture is implemented in the Test data set and result is listed below for a single search. The keyword chosen is 'cluster'. The result is as shown, for each feature selection criteria for the best class selected is listed and among those best retrieved class, best document is retrieved. Also it shows the mean value of the iterations and the number of documents in each cluster.

Search and retrieval have considered time as a core dimension. In a dynamic environment, every single ingredient of the retrieval pipeline needs to be on-line modifiable, without lengthy test-train-deploy-update cycles. To make this happen systematically, we need a framework to talk about time and to test system performance vis-à-vis time.

The summary of the result is shown below and time taken for single run is depicted in Table-2.

Selection function	Best Class	k-means value	Time (ms)
IG	PPT	7.270	109
CHI-SQUARE	PPT	7.270	47
CC	JS	11.69	47
GSS	PDF	3.634	16

Table-2 MFCC Search summary

An information system is affected by time in many ways. The information processes changes continuously both in context and from the world that information references evolves, and information needs and usage scenarios change and evolve. In a big data context, modeling the character, content and evolution of a steadily changing immense information stream requires a perspective of information as something dynamic over time, not as something constant to be extracted.

The test data is verified with Hard k-means MFCC. The result is shown Fig-4; the Hard k-means clusters the classification into two clusters according to search word or the query.

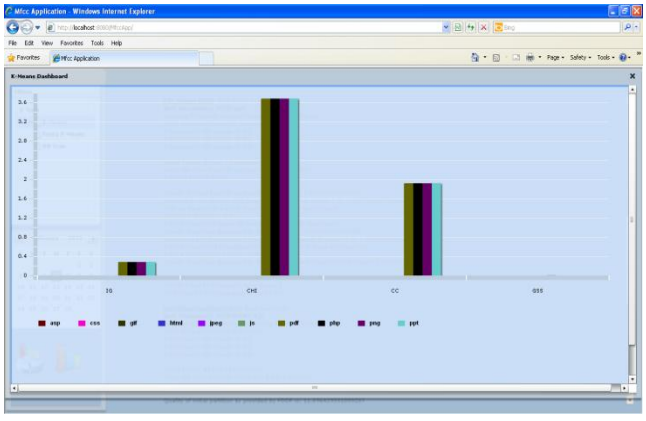

Fig-4: MFCC & Hard k-means algorithm in Searching Strategy

IV. CONCLUSION

MFCC exploits the different types of feature classes to perform web document clustering. It has been implemented in search engine technology to improve the rank results. The co-selection among other feature space, and intermediate clustering results in fusion function. So that the database index is fresh and always takes the entire web pages for clustering. Finally, the usage of MFCC in IR searching architecture reduces the noisy data. It removes the challenge of search engine, the freshness of newly loaded pages. The future scope of the architecture frame is put to test and continued for other data sets than textual. The future, we plan to test MFCC on more data sets. Also try to find better strategies for the co-selection of features having different efficiency in clustering. In addition, the co-selection idea will be tested using some clustering algorithms other than hard k-means, like soft clustering, hierarchical clustering, and density-based clustering.

REFERENCES

[1] Ed. Green grass, "Information Retrieval: A survey"; 2000.

[2] Report from SWIR 2012; “Frontiers, Challenges, and Opportunities for IR”; ACM SIGIR forum vol. 46, No.1, June 2012.

[3] Sew Staff, "How search engines work", 2007.

[4] Han & et al., "Multi type feature co-selection for clustering for web documentation", IEEE transaction on knowledge engineering, June 2006.

[5] K.Parimala, Dr.V.Palanisamy, “Enhanced Performance of Search Engine with Multitype Feature Co-Selection for Clustering Algorithm”, International Journal of Computer Applications (0975 - 8887) Volume-53- No. 7, September2012.